HW 6

Zoe Werner

11/18/2024

What is the difference between gradient descent and *stochastic* gradient descent as discussed in class? (*You need not give full details of each algorithm. Instead you can describe what each does and provide the update step for each. Make sure that in providing the update step for each algorithm you emphasize what is different and why.*)

Student Input

Gradient descent is the vector direction of the steepest descent and uses a learning rate variable – alpha – to determine the minimum value. Stochastic gradient descent also calculates the descent, but uses a random subset of data, which increases variability in calculating gradience and prevents getting stuck in local extremes. Gradient Descent: $\theta_{i+1} = \theta_i - \alpha \nabla f(\theta_i, X, Y)$ Stochastic Gradient Descent: $\theta_{i+1} = \theta_i \alpha \nabla f(\theta_i, X_I, Y_I)$ The (X_I, Y_I) variables in the stochastic variation differ from the (X,Y) variables in gradient descent. The (X_I, Y_I) variables are the random subset and the I is an indexing variable, which differs from i used elsewhere. Both of these variables are used in stochastic gradient descent, but not in regular gradient descent.

Consider the `FedAve` algorithm. In its most compact form we said the update step is $\omega_{t+1} = \omega_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} \nabla F_k(\omega_t)$. However, we also emphasized a more intuitive, yet equivalent, formulation given by $\omega_{t+1}^k = \omega_t - \eta \nabla F_k(\omega_t); w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$.

Prove that these two formulations are equivalent.
(*Hint: show that if you place ω_{t+1}^k from the first equation (of the second formulation) into the second equation (of the second formulation), this second formulation will reduce to exactly the first formulation.*)

Student Input

1. $\omega_{t+1}^k = \omega_t - \eta \nabla F_k(\omega_t); w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$
2. $w_{t+1} = \sum_{k=1}^{k} \frac{nk}{n} (w_t - \eta \nabla F_k(w_t))$ Plug w_{t+1}^k into the second section of the forumation.
3. $w_{t+1} = \sum_{k=1}^{k} \frac{nk}{n} w_t - \frac{nk}{n} \eta \nabla F_k(w_t)$ Distribute $/fracnkn$ to w_t and $\eta \nabla F_k(w_t)$.
4. $w_{t+1} = \sum_{k=1}^{k} \frac{nk}{n} w_t - \sum_{k=1}^{k} \frac{nk}{n} \eta \nabla F_k(w_t)$ Distribute the summation notation to both sections (sections stated above).
5. $w_{t+1} = w_i \sum_{k=1}^{k} \frac{nk}{n} - \eta \sum_{k=1}^{k} \frac{nk}{n} \nabla F_k(w_t)$ Move w_i to the front of the first section.
6. $w_{t+1} = w_t - \eta \sum_{k=1}^{k} \frac{nk}{n} \nabla F_k(w_t)$ Since $nk = n$ and $\frac{nk}{n} = 1$, it cancels out and the first section is left with w_i. This is the simplified, more compact version of the FedAve formation.

Now give a brief explanation as to why the second formulation is more intuitive. That is, you should be able to explain broadly what this update is doing.

Student Input

The second formulation of the FedAve algorithm is more intuitive becuase its broken into two parts, allowing you to see the entire process outlined. The first equation shows the update for each of the local K clients and the second equation shows the weighted averaging for the global level. You can explicitly see the steps of local model averaging and global updates. Since the process is broken up and explicitly shows the steps of FedAve, it is the more intuitive option.

Prove that randomized-response differential privacy is ϵ-differentially private.

Student Input

ϵ-differential privacy is the privacy for $\in > 0$ if D_1, D_2 differ in exactly one element. It basically states that the difference between testing and training data is minute.

ϵ-differential privacy: $\frac{P[A(D_1) \in S]}{P[A(D_2) \in S]} \leq e^{\in}$

Randomized response differential privacy is ϵ-differentially private using the example from class of flipping a coin and using the results to answer the question "have you ever cheated?" The student will flip a coin and not let the surveyor see the result. If it is heads, they will flip the coin again, but tell the truth regardless of the answer. If the coin lands on tails the first time, flip it again and if it is heads, say yes, and if it is tails, say no, regardless of the truth.

Assume that D and S $\in$ {Yes, No} and S = Yes

1. $\frac{P[A(Yes)=Yes]}{P[A(No)=Yes]} = \frac{P[Output=Yes|Input=Yes]}{P[Output=Yes|Input=No]} = \frac{3/4}{1/4} = 3 = e^{ln(3)}$

The above demonstrates that there are 3 chances of a student replying yes if the truth is yes and there is 1 change that a student will reply no if yes is the truth.

2. $\frac{P[A(No)=Yes]}{P[A(Yes)=Yes]} = \frac{1}{3}$

During this step, we flipped the equation (swapped denominator and numerator) to get the other bound. Since it is $\frac{1}{3}$, we will use 3 because it includes both bounds.

3. $\frac{P[A(D_1) \in S]}{P[A(D_2) \in S]} \leq 3 = e^{ln(3)}$

The above is the predicted probability for the testing data based on the previous steps.

4.$RR = ln(3) - DP$

Define the harm principle. Then, discuss whether the harm principle is *currently* applicable to machine learning models. (*Hint: recall our discussions in the moral philosophy primer as to what grounds agency. You should in effect be arguing whether ML models have achieved agency enough to limit the autonomy of the users of said algorithms.*)

Student Input

The harm principle was developed by J.S. Mill and states that personal autonomy is restricted when using autonomy would result in objective moral harm. According to the principle of Deontology, an act is not permissible if it treats moral agents as a means to an end. Moral agents are defined as humans so following this precedent, machine learning models do not have agency because they are not human. However, the line continues to blur as ML models become more aware and capable of understanding human emotions. The question becomes: if something is capable of human-level sentience and consciousness, should it fall into the category of human? Other definitions of moral agency describe it as an entity that can make independent decisions and are responsible for maintaining morality.This again arises the question of the full capabilities of ML models. Morality is rooted in both philosophical frameworks and personal beliefs. Since a ML model does not (yet) have the capacity to form thoughts independently from their algorithm and/or the use of the internet, I still claim that they do not have personal agency. I would also conclude that current ML models are not moral agents because understanding human emotion is not the same as feeling them. If we reach a technological point in the future where scientists are confident that ML models have achieved emotional capacity, sentience, and consciousness comparable to humans, then they should be considered moral agents and treated accordingly. However, for the time being, they are not technologically advanced enough to be classifed as moral agents. At this moment in time, the harm principle should not apply not machine learning models since they have not achieved moral agency.